维普资讯 http://www.cqvip.com

基于改进的 PCM 支持向量描述多类分类器*)

张 永[1,2] 迟忠先[2] 谢福鼎[1]

(辽宁师范大学计算机系 大连 116029)[1] (大连理工大学计算机科学与工程系 大连 116024)[2]

摘 要 基于支持向量数据描述和改进的可能性 c-均值聚类算法,提出了一种模糊的多类分类学习机。首先通过一个改进的 PCM 算法来计算每个样本对于每类的权值矩阵,该权值也反映了该样本对某类的重要程度;然后将该权值矩阵应用到支持向量数据描述方法中,并对样本进行训练;最后给出了一个针对多类分类的分类规则(函数),并从理论上证明该分类规则满足贝叶斯优化决策理论。通过对比实验分析,本文提出的算法在分类精度和训练时间上都有较大的改善。

关键词 支持向量数据描述,可能性 c-均值聚类,最小包围球,分类器,支持向量机

Support Vector Description Multi-class Classifier Based on Improved PCM

ZHANG Yong[1,2] CHI Zhong-xian[2] XIE Fu-ding[1]

(Department of Computer, Liaoning Normal University, Dalian 116029, China)[1]

(Department of Computer Science and Engineering, Dalian University of Technology, Dalian 116024, China)[2]

Abstract In this paper, a novel fuzzy classifier for multi-classification problems, based on Support Vector Data Description (SVDD) and improved PCM, is proposed. The proposed method is the robust version of SVDD by assigning a weight to each data point, which represents fuzzy membership degree of the cluster computed by the improved PCM method. Accordingly, this paper presents the multi classification algorithm based on the robust weighted SVDD, and gives the simple classification rule. Experimental results show that the proposed method can reduce the effect of outliers and yield higher classification rate.

Keywords Support vector data description, Possibilistic c-means clustering, Minimum enclosing sphere, Classifier, SVM

1 引言

支持向量机本身是用来解决二类分类问题的,当将其扩展到多类分类问题中时,存在一个不可分区域。常用的方法之一是将一个多类问题转化为多个二类问题,比如 one-against-one, ones-against-rest 以及 DAGSVMs[1,2] 等方法。另外一个解决思路是直接构造一个多类分类器,比如 Weston 和 Watkins[3] 提出的 k-SVM,在此基础上 Zhang 等人[4] 提出了模糊补偿多类分类器方法。最近,一些学者[5,6] 基于最小包围球的思想,也提出了一些新的方法来解决多类分类问题。

所谓最小包围球,最初由 Schölkopf 等人[7,8] 提出,用来对支持向量分类器中的 VC 维进行估计。受 Schölkopf 等人的启发,Tax 和 Duin[9,10] 又将最小包围球的概念应用到数据域描述问题中,并称为支持向量数据描述(Support Vector Data Description, SVDD)。从计算角度看,基于最小包围球的模式分类有一大优势:样本往往仅被训练一次。当将其应用到多类分类问题中时,这一优势尤为明显,因为基于某一类中的样本可以被单独训练,而且仅需训练一次即可。相对于 one-against-one 或 one-against-rest 方法[1,2],基于最小包围球的多类分类能明显减少训练时间。基于最小包围球的思想,Zhu 等人[5] 提出了一个多类分类算法,并与支持向量机方法做了实验对比;在此基础上,Wang 等人[6] 和 Lee 等人[11] 也分别提出了基于数据域描述的多类分类方法。

由于支持向量数据描述方法往往只需要一类样本数据就可以进行分类,因此具有计算速度快、鲁棒性强等优点,且具有较好的泛化能力,已在许多异常值检测问题或 1-类分类问题中得到了成功的应用[18-20]。尽管该方法具有较好的推广能力,但由于在构造最优超球面时所有的样本具有同等的地位,因此当训练样本中含有一些噪声或野值样本时,这些含有"异常"信息的样本在特征空间中常常位于超球面附近,导致获得的超球面不是真正的最优超球面。针对这种情况,Song 等人[12,13] 提出了均值算法来降低异常点对分类结果的影响,但附加的参数导致了优化的困难。Lin 等学者提出了模糊支持向量机方法(FSVM)[14,15],将模糊技术应用于支持向量机中,对不同的样本采用不同的惩罚权系数,使得在构造目标函数时,不同的样本有不同的贡献,对含有噪声或野值的样本赋予较小的权值,从而达到消除噪声与野值样本影响的目的。然而该方法中样本隶属度值的计算并没有一致的方法。受 Lin 工作的启发,本文将模糊思想引入到支持向量数据描述中,并给出了一种基于改进的 PCM 算法来计算模糊隶属值的方法。

在本文中,我们首先通过一个改进的 PCM 算法来计算每个样本对于每类的权值矩阵,该权值也反映了该样本对某类的重要程度;然后将该权值矩阵应用到支持向量数据描述方法中,并对样本进行训练;最后给出了一个针对多类分类的分类规则(函数),并从理论上证明该分类规则满足贝叶斯优

*)国家科技型中小企业技术创新基金(05C26212120357)。张 永 博士,讲师,主要研究领域为机器学习、智能计算、数据挖掘;迟忠先 教授,博士生导师,主要研究领域为机器学习、数据挖掘、数据仓库、软件工程;谢福鼎 博士,教授,主要研究领域为人工智能、知识发现。

维普资讯 http://www.cqvip.com

化决策理论。

2 数据域描述问题

数据域的描述问题是模式识别中继分类和回归问题之后又一个需解决的问题。其基本任务是对训练样本所在的类进行描述，并且能拒绝来自其它类的数据，数据域描述可用于异常值识别和新模式的发现。随着支持向量机的发展以及特殊分类问题的出现，支持向量机也已经开始寻求专门解决这类问题的方法。Schölkopf 等人[7,8]提出了最小包围球，最初用来对支持向量分类器中的 VC 维进行估计；受 Schölkopf 等人的启发，Tax 和 Duin[9,10]又将最小包围球的概念推广到数据域描述问题中，并称之为支持向量数据描述。

SVDD 方法的主要思想是：首先把训练数据通过非线性变换映射到一个高维特征空间，然后在此特征空间中去寻找尽可能多的包围这些映射数据的最小球体，称之为最小包围球。让目标样本点尽可能被包围在最小包围球体中，而非目标样本点尽可能不被包含在最小包围球体中，从而实现两类之间的划分。这种方法往往只需要一类样本数据即可进行分类且具有计算速度快、鲁棒性强、可有效处理小样本数据等优点。

设有给定的数据集$\{x_i, i=1,2,\cdots,l\}\subset\mathcal{R}^n$，最小包围球记为 S，其中心为 a，半径为 R。最小包围球 S 可通过求解下面带约束的二次优化问题得到：

$$\min_{a,R} \quad R^2+C\sum_i\xi_i \tag{1}$$

$$\| \phi(x_i)-a \|^2\leqslant R^2+\xi_i$$

$$\xi_i\geqslant 0, \quad \text{for} \quad i=1,2,\cdots,l \tag{2}$$

其中：a 是最小包围球 S 的中心；$C>0$ 是一个惩罚常量，用来在最小包围球的大小和可能落在球体外的样本数量之间进行平衡；$\xi_i\geqslant 0$ 是松弛变量；ϕ 是一个非线性映射。

通常，运用 Lagrange 乘子法来求解该问题，可得其对偶形式：

$$\max \quad W=\sum_i\alpha_iK(x_i,x_i)-\sum_{i,j}\alpha_i\alpha_jK(x_i,x_j)$$

$$s.t. \quad 0\leqslant\alpha_i\leqslant C, \quad \sum_i\alpha_i=1, \quad j=1,2,\cdots,l \tag{3}$$

其中，核函数 $K(x_i,x_j)=\phi(x_i)\cdot\phi(x_j)$ 必须满足 Mercer 条件。仅仅那些使得 $\alpha_i>0$ 的样本点才被称为支持向量(SVs)。

为了测试一个新的样本点 x 是否属于该类，只需计算该样本点到球体中心的距离是否大于球体的半径 R：

$$\| \phi(x)-a \|^2=K(x,x)-2\sum_i\alpha_iK(x_i,x)+\sum_{i,j}\alpha_i\alpha_jK(x_i,x_j) \tag{4}$$

很显然，式中 R^2 表示在边界上的任意一个支持向量 x_i 到球体中心 a 的距离。

3 改进的 PCM 算法

受 Lin 工作的启发，本文将模糊思想引入到支持向量数据描述中。考虑到可能性 c-均值聚类算法(possibilistic c-means clustering, PCM)[16,17]对噪声和异常点数据的较好处理能力，我们对原始的 PCM 算法做了一些改进，用来计算模糊隶属值。

可能性 c-均值聚类算法(PCM)最初由 Krishnapuram 等人提出[16]，然后又在文献[17]做了进一步的改进。该算法克服了模糊 c-均值聚类算法(FCM)中的一些缺点，对处理含有噪声和异常点的训练集有较好的效果。其基本思想是放松模糊 c-均值聚类算法中的约束条件，以概率的形式来对样本进行聚类，最终是对下面的目标函数最小化[16]：

$$J(U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^m\| x_k-v_i \|^2+\sum_{i=1}^{c}\eta_i\sum_{k=1}^{n}(1-u_{ik})^m \tag{5}$$

其中：c 是事先确定的簇数，本文中选定其为待分类训练样本的类别数；n 训练样本的数量；$u_{ik}\in[0,1]$表示样本 x_k 隶属于第 i 类的程度；$m\in[1,\infty)$是模糊加权指数，用来控制聚类的模糊程度，也称为平滑参数，它对聚类的模糊程度有重要的调节作用；$V=\{v_1,v_2,\cdots,v_c\}$是聚类中心或模式原型($v_i\in R^p$)；参数 η_i 为适当的正值，可以理解为与聚类的宽度有关。

为了将 PCM 算法应用到本文提出的加权支持向量描述多类分类器中，我们将原始的 PCM 算法扩展到核空间中，从而用在核空间中生成的隶属度值代替了在原始输入空间中生成的隶属度值。

定义一个非线性映射 $\phi: x\to\phi(x)\in F$，其中 $x\in X$。X 表示数据空间，F 表示具有更高维的映射特征空间。基于核的 PCM 算法就是使得下面的目标函数最小化：

$$\min J(U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^m\| \phi(x_k)-\phi(v_i) \|^2+\sum_{i=1}^{c}\eta_i\sum_{k=1}^{n}(1-u_{ik})^m \tag{6}$$

其中：$\| \phi(x_k)-\phi(v_i) \|^2=K(x_k,x_k)+K(v_i,v_i)-2K(x_k,v_i)$。

通常情况下，式(6)采用高斯核，即 $K(x_i,x_j)=\exp(-\| x_i-x_j \|^2/\sigma^2)$。显然此时有 $K(x,x)=1$，$\| \phi(x_k)-\phi(v_i) \|^2=2-2K(x_k,v_i)$。相应地，式(6)可简化为如下目标函数：

$$\min J(U,V)=2\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^m(1-K(x_k,v_i))+\sum_{i=1}^{c}\eta_i\sum_{k=1}^{n}(1-u_{ik})^m \tag{7}$$

引用文献[16,17]的建议，η_i 可通过计算第 i 类的类内加权均值距离得到，即

$$\eta_i=K\frac{\sum_{k=1}^{n}u_{ik}^m\| \phi(x_k)-\phi(v_i) \|^2}{\sum_{k=1}^{n}u_{ik}^m}=K\frac{2\sum_{k=1}^{n}u_{ik}^m(1-K(x_k,v_i))}{\sum_{k=1}^{n}u_{ik}^m} \tag{8}$$

一般情况下，K 取值为 1，本文中我们也取 K=1。

在式(5)的最小化迭代过程中，可通过 $J(U,V)$ 分别对 u_{ik} 和 v_i 求偏导，并令其值为 0，从而得到下面两个式子来更新隶属度值 u_{ik} 和聚类中心 v_i[16]：

$$u_{ik}=\frac{1}{1+(\| \phi(x_k)-\phi(v_i) \|^2/\eta_i)^{1/(m-1)}}=\frac{1}{1+(2(1-K(x_k,v_i))/\eta_i)^{1/(m-1)}} \tag{9}$$

$$v_i=\frac{\sum_{k=1}^{n}u_{ik}^m\phi(x_k)}{\sum_{k=1}^{n}u_{ik}^m} \tag{10}$$

基于核的 PCM 算法就是一个通过式(9)和式(10)不断更新隶属度值 u_{ik} 和聚类中心 v_i 的迭代过程，直到停止条件满足才终止迭代。

经过上述的迭代过程进行数据分区后，噪声和异常点的模糊隶属度值将很小，甚至为 0，这样这些样本点对最终分类决策面的影响将降至很低。另外，为了减少下面提到的加权支持向量描述多类分类算法的训练时间，我们对所有的模糊隶属度值 u_{ik} 设定了一个阈值 σ_t：如果 $u_{ik}>\sigma_t$，表明训练样本 x_k 对第 i 类产生了一定的影响；否则，训练样本 x_k 对第 i 类没有多大的影响，我们近似地认为影响为 0，故相应地修改模糊隶属度值 $u_{ik}=0$。

维普资讯 http://www.cqvip.com

4 基于改进的 PCM 加权 SVDD

4.1 加权 SVDD 方法

给定一个训练数据集 $\{(x_i,m_i),i=1,2,\cdots,l\}$，其中 $x_i\subset\mathcal{R}^n$ 为训练样本，m_i 是样本 x_i 的权值，该值是通过上述改进的 PCM 算法计算而得的。这样加权支持向量数据描述方法可通过下面的优化问题来求解：

$$\min_{a,R} R^2+C\sum_i m_i\xi_i \tag{11}$$

约束条件为

$$\|\phi(x_i)-a\|^2\leqslant R^2+\xi_i$$

$$\xi_i\geqslant 0,\text{for } i=1,2,\cdots,l \tag{12}$$

其中：a 是最小包围球 S 的中心；R 为 S 的球半径；$\xi_i\geqslant 0$ 是松弛变量；m_i 是样本 x_i 的权值；$C>0$ 是一个惩罚常量，用来在最小包围球的大小和可能落在球体外的样本数量之间进行平衡。

为了求解带约束条件的优化问题式(11)，我们可以把其转换为相应的 Lagrange 函数，并表示为

$$L=R^2-\sum_i(R^2+\xi_i-\|\phi(x_i)-a\|^2)\alpha_i-\sum_i\beta_i\xi_i+C\sum_i m_i\xi_i \tag{13}$$

其中，$\alpha_i\geqslant 0,\beta_i\geqslant 0$ 为 Lagrange 系数。

求解式(13)的最小值，可以令该泛函分别对 R,c,ξ_i 求偏导，并令它们的值都等于 0，得到

$$\frac{\partial L}{\partial R}=2R(1-\sum_i\alpha_i)=0 \tag{14}$$

$$\frac{\partial L}{\partial c}=2\sum_i\alpha_i(\phi(x_i)-a)=0 \tag{15}$$

$$\frac{\partial L}{\partial \xi_i}=Cm_i-\alpha_i-\beta_i=0 \tag{16}$$

从式(14)中，有 $\sum_i\alpha_i=1$；从式(15)中，有 $a=\sum_i\alpha_i\phi(x_i)$。

将式(14)、式(15)及式(16)代入式(13)，则基于加权的支持向量数据描述的最优化问题转化为下面的对偶问题：

$$\max W=\sum_i\alpha_i K(x_i,x_i)-\sum_{i,j}\alpha_i\alpha_j K(x_i,x_j)$$

$$s.t.\ 0\leqslant\alpha_i\leqslant Cm_i,\sum_i\alpha_i=1,j=1,2,\cdots,l \tag{17}$$

通过对式(17)二次规划问题的求解，可得各样本对应的 Lagrange 系数 α_i，最小包围球 S 的中心 a 可表示为 x_i 的加权平均 $a=\sum_i\alpha_i x_i$。

根据 KKT 条件，有

$$\alpha_i=0\Rightarrow\|\phi(x_i)-a\|^2<R^2 \text{ and } \xi_i=0$$

$$0<\alpha_i<Cm_i\Rightarrow\|\phi(x_i)-a\|^2=R^2 \text{ and } \xi_i=0$$

$$\alpha_i=Cm_i\Rightarrow\|\phi(x_i)-a\|^2\geqslant R^2 \text{ and } \xi_i\geqslant 0 \tag{18}$$

由式(17)解出的 α_i 中，只有部分(通常是少部分)大于 0，其它均为 0。对于那些 $\alpha_i\geqslant 0$ 所对应的样本 x_i，才被称为支持向量。从式(18)中可以看出，支持向量可分为两种情况：其一是当 $\alpha_i=Cm_i$ 时，$\|\phi(x_i)-a\|^2\geqslant R^2$，此时对应的样本点 x_i 在超球面外部(也可能在球面上)；其二是当 $0<\alpha_i<Cm_i$ 时，$\|\phi(x_i)-a\|^2=R^2$，并且 $\xi_i=0$，此时对应的样本点 x_i 位于超球面上。

标准支持向量数据描述方法与加权支持向量数据描述方法的区别在于对偶问题中的 Lagrange 乘子 α_i 的上界不同。在标准支持向量数据描述中，Lagrange 乘子 α_i 的上界为常数 C ($0<\alpha_i<C$)，而在加权支持向量数据描述中，Lagrange 乘子 α_i 的上界为 Cm_i ($0<\alpha_i<Cm_i$)。

4.2 提出的算法描述

结合前面的分析，这一部分我们给出加权支持向量数据描述针对多类分类问题的相关方法。设有训练集 (x_1,y_1)，$(x_2,y_2),\cdots,(x_l,y_l)$，其中 $l\in R$ 表示训练样本的数量，$x_i\in\mathcal{R}^n$ 表示一个输入模式，$y_i\in R=\{w_1,w_2,\cdots,w_k\}$ 表示训练集的输出类别。我们提出的方法的主要思想是，充分利用模糊支持向量数据域描述所生成的信息去估计每个类中数据的分布情况，然后通过 Bayesian 决策规则来分类新样本。

详细的算法描述如下。

Step1 (计算权值 u_{ik}) 设有阈值 σ_t，运用改进的基于核的 PCM 方法来计算每个样本 x_k 相对于每一类 i 的权值 u_{ik}，从而可以得到一个权值矩阵 $U=\{u_{ik}\}$。并且约定：如果 $u_{ik}\leqslant\sigma_t$，则 $u_{ik}=0$。

Step2 (数据分区) 根据第一步计算出来的权值 u_{ik}，将给定的训练样本集分成 c 个子集 $\{D_p\}_{p=1}^c$，其中，这里的 c 表示训练样本的类别数。比如，第 p 个子集 D_p，包含了 l_p 个如下所示的样本：

$$D_p=\{(x_{k_1},w_p),\cdots,(x_{i_{k_p}},w_p)\} \tag{19}$$

其中，$p=1,2,\cdots,c;u_{k_dp}\geqslant\sigma_t$，对于任意的 $d=1,\cdots,l_p$。

Step3(针对每类数据集的加权 SVDD) 对于任意一个数据集 D_p，利用上述的方法建立基于改进 PCM 的加权 SVDD 分类器。特别地，为叙述方便，假设对偶问题(17)的解为 α_i^*，$i=i_1,\cdots,i_{l_p}$，并设 $J_p\subset\{1,\cdots,l_p\}$ 为非零 α_i^* 的下标集合。则数据集 D_p 所对应的 SVDD 分类函数为

$$f_p(x)=K(x,x)-2\sum_{i\in J_p}\alpha_i^*K(x_i,x)+\sum_{i,j\in J_p}\alpha_i^*\alpha_j^*K(x_i,x_j) \tag{20}$$

根据前面的分析，我们知道要判断一个新样本 x 是否属于某类，只需计算该样本 x 到类中心 a 的距离是否小于最小包围球的球半径。如果是，则该样本属于此类；否则不属于此类。

Step4(构造分类规则) 对于任意一个新样本 x，我们给定如下的分类规则

$$F(x)=\arg\max_p\frac{l_p}{l}((R_p^*)^2-f_p(x)) \tag{21}$$

其中，$p=\{w_1,w_2,\cdots,w_k\}$，对于第 p 类中任意一个支持向量 x_i 有 $(R_p^*)^2=f_p(x_i)$；$l\in R$ 表示训练样本的数量，$l_p\in R$ 表示第 p 类训练样本的数量。

该分类规则满足贝叶斯优化决策理论，在最小化平均误差概率的情况下，将一个新样本尽可能的分到了具有最大后验概率的类别中，下面的章节中对此做进一步的解释和说明。

5 算法理论证明

5.1 贝叶斯决策理论

贝叶斯决策论是解决模式分类问题的一种基本统计途径，其基本思想比较简单。为最小化总风险，总是选择那些能够最小化条件风险 $R(\alpha|x)$ 的行为。尤其是，为了最小化分类问题中的误差概率，总是选择那些使后验概率 $p(\omega_i|x)$ 最大的类别。贝叶斯公式允许我们通过先验概率 $p(\omega_i)$ 和条件密度概率 $p(x|\omega_i)$ 来计算后验概率。如果对在模式 ω_i 中所做的误分的惩罚与模式 ω_j 的不同，那么在做出判决行为之前须先根据惩罚函数对后验概率加权。

设有 $\{\omega_1,\omega_2,\cdots,\omega_c\}$ 表示有限的 c 个类别集，$\{\alpha_1,\alpha_2,\cdots,\alpha_a\}$ 表示有限的 a 种可能采取的行为集，风险函数 $\lambda(\alpha_i|\omega_j)$ 描述类别状态为 ω_j 时采取行动 α_i 的风险。设特征向量 x 表示

维普资讯 http://www.cqvip.com

一个 d 维随机变量。设 $p(x|\omega_i)$ 表示 x 的类条件概率密度(即类别状态为 ω_i 时的 x 的概率密度函数,又称之为状态条件概率密度函数),$p(\omega_i)$ 表示类别处于状态 ω_i 时的先验概率,则后验概率 $p(\omega_i|x)$ 可通过贝叶斯公式计算得到:

$$p(\omega_i|x)=\frac{p(x|\omega_i)p(\omega_i)}{p(x)} \tag{22}$$

而证据(evidence)因子 $p(x)$ 可表示为:

$$p(x)=\sum_{j=1}^{c}p(x|\omega_j)p(\omega_j) \tag{23}$$

从而有

$$p(\omega_i|x)=\frac{p(x|\omega_i)p(\omega_i)}{p(x)}=\frac{p(x|\omega_i)p(\omega_i)}{\sum_{j=1}^{c}p(x|\omega_j)p(\omega_j)} \tag{24}$$

最小化风险的贝叶斯决策规则要求选择一种能使条件风险最小化的行为。因此,为了最小化平均误差概率,我们需要选取 i 使得后验概率 $p(\omega_i|x)$ 最大。即基于最小化误差概率,有

对任给 $i\neq j$,如果 $p(\omega_i|x)>p(\omega_j|x)$,则判决为 ω_i。

5.2 提出算法的理论说明

基于上述对贝叶斯决策理论的分析,我们很容易说明我们提出算法中的分类规则是一种贝叶斯决策分类规则。

定理 1 算法所提供的分类规则式(21)满足贝叶斯决策分类规则。

证明:一个贝叶斯分类器的结果可由条件概率密度 $p(x|\omega_i)$ 和先验概率 $p(\omega_i)$ 来决定。其中先验概率 $p(\omega_i)$ 的估计较为容易,常用某类样本数量在整个训练样本数据中的比例来表示。记为

$$p(\omega_i)=\frac{l_i}{l} \tag{25}$$

其中,l,l_i 与算法第 4 步中的变量 l,l_p 含义相同。$l\in R$ 表示训练样本的总数量,$l_i\in R$ 表示第 i 类训练样本的数量。

对于条件概率密度 $p(x|\omega_i)$,可以表示成伪概率密度 $\hat{p}(x|\omega_i)$,记为

$$\hat{p}(x|\omega_i)=(R_i^*)^2-f_i(x) \tag{26}$$

这里 $i=\{w_1,w_2,\cdots,w_k\}$ 表示类别,对于第 i 类中任意一个支持向量 x_j 有 $(R_i^*)^2=f_p(x_j)$。

从而有后验概率

$$p(\omega_i|x)=\frac{\hat{p}(x|\omega_i)p(\omega_i)}{p(x)}=\frac{\frac{l_i}{l}((R_i^*)^2-f_i(x))}{p(x)} \tag{27}$$

注意后验概率主要由先验概率和条件概率密度的乘积来决定,证据因子 $p(x)$ 可仅仅看成是一个标量因子,以保证各类别的后验概率总和为 1,从而满足概率条件。换言之,在比较不同 ω_i 值的后验概率时,分母 $p(x)$ 总是常数,因此可以忽略。

根据贝叶斯决策理论,一个新的样本点因该被分到具有最大后验概率的类别中,即有分类规则

$$F(x)=\arg\max_i p(\omega_i|x) \tag{28}$$

显然,根据上面的分析,式(28)与下面的分类规则等价:

$$F(x)=\arg\max_i \frac{l_i}{l}((R_i^*)^2-f_i(x)) \tag{29}$$

式(29)与算法第 4 步中的分类规则(21)等价。证明完毕。

6 实验分析

为了验证我们提出的加权支持向量数据描述算法的有效性,从 UCI 机器学习资料库[21]中选取了一些标准数据集来进行测试。共包括以下 6 个数据集:glass,ionosphere,onar,pima-diabetes,iris 和 vehicle。这些数据集的相关参数情况详见表 1。其中,#pts 表示数据集中样本数量;#att 表示属性的个数(向量维数);#class 表示类别数。

对于每个数据集,我们将本文提出的算法分别与文献[5,6]中的基于球的分类器算法和 SVM 算法进行了实验比较,实验采用高斯核和 5-重交叉验证方法。为便于实验比较,在交叉验证过程中,我们对每个算法都采用了相同的训练集和测试集,并对参数 C 和 σ 运用了格子搜索方法来选取。相应地,对于算法中在 Step1 中提出的阈值 σ_t,设定其值为 0.1。阈值 σ_t 将在一定程度上影响算法的运行时间。

实验首先对数据进行了统一的归一化处理。选取高斯函数作为核函数,参数 σ^2 的选择范围是 $[2^{-15},2^{-13},\cdots,2^1,2^3]$,参数 C 的选择范围是 $[2^{-3},2^{-1},\cdots,2^{13},2^{15}]$,这样共有 $10\times 10=100$ 种参数组合方式。实验是在 Celeron 1.5GHz 处理器、512MB RAM 的计算机上进行的。

表 2 对比了本文提出的加权 SVDD 分类方法与文献[5,6]中的基于球的分类器算法、标准 SVM 算法对各数据集的错误分类率和训练时间。

表 1 标准数据集

数据集	#pts	#att	#class
Ionosphere	351	34	2
Sonar	208	60	2
Pima-diabetes	768	8	2
Iris	150	4	3
Vehicle	846	18	4

表 2 各种方法分类精度及训练时间比较

Data Set	SVM		Sphere-based classifier		Weighted SVDD ($\sigma_t=0.1$)	
	分类错误率	训练时间	分类错误率	训练时间	分类错误率	训练时间
Glass	27.57	7.56	28.54	3.48	27.51	5.85
Ionosphere	5.12	12.47	4.84	8.56	4.68	11.46
Sonar	10.62	8.54	15.12	5.37	11.38	8.01
Pima-diabetes	27.48	32.35	27.00	19.59	26.82	23.72
Iris	2.67	0.23	2.72	0.15	2.67	1.35
Vehicle	12.64	48.43	12.60	32.67	12.60	38.75
Iris*	5.42	0.32	5.64	0.21	3.89	1.39
Vehicle*	19.86	50.78	19.32	34.70	17.58	39.93

从表 2 中可以看出,除了数据集 Sonar 外,相对于标准支持向量机而言,我们提出的加权支持向量数据描述多类分类方法获得了更好的分类结果。同时与文献[5,6]中用到的基于最小包围球方法相比较,在这 6 个数据集中,本文提出的算法都获得了更好的精度。

另外,为了测试算法对噪声的敏感程度,我们对于数据集 iris 和 vehicle,随机地增加了 10%的"异常"数据。表 2 中的后两行(标记为 iris* 和 vehicle*)显示了相应的试验结果。从结果中,我们发现本文提出的算法误分率最低,分别为 3.89%和 17.58%。相对其它两种方法而言,分类效果有明显的改善。

为了更直观地对数据进行比较分析,我们给出了训练时间的对比图例,如图 1 所示。为了简化图表,图中 X 坐标轴上分别用编号 1 至 8 表示表 2 中的 8 个数据集。从时间角度来分析,由于最小包围球方法每个训练样本仅训练一次,因此

维普资讯 http://www.cqvip.com

训练时间最低，但获得的分类效果最差。本文提出的算法由于采用了改进的PCM算法来计算样本的隶属度值，使得部分样本(实际上是很少部分样本)可能会被训练多次，导致了时间开销上比最小包围球方法稍大。相对于标准SVM而言，在数据集Sonar上，时间开销基本相当，训练时间差不多，但在其它几个数据集上，时间优势还是比较明显的。

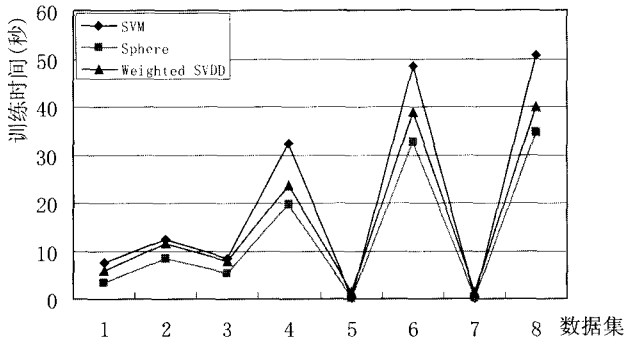

图1 训练时间对比图

结束语 针对1-类分类问题中的支持向量数据描述(SVDD)，根据对其特点的相关分析，本文将其成功地应用到多类分类问题中。另外，为了更好地处理训练数据中的噪声和异常点，在算法中引入了模糊技术，并采用一种改进的PCM方法来计算样本的模糊隶属度值，从而提出了一种基于改进PCM方法的加权支持向量数据描述多类分类算法。经理论证明，该算法所提出的分类决策函数满足贝叶斯优化决策理论。最后将该算法与其它的方法进行了实验比较，得到了较满意的结果。

参 考 文 献

[1] Platt J C, Cristianini N, Shawe-Taylor J. Large margin DAGs for multiclass classification. Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2000: 547-553

[2] Hsu C W, Lin C J. A comparison of methods for multi-class support vector machines. IEEE Trans. on Neural Networks, 2002, 13(2): 415-425

[3] Weston J, Watkins C. Multi-class support vector machines. Technical Report, SD2TR298204. Department of Computer Science, Royal Holloway University of London, 1998

[4] Zhang Y, Chi Z X, Liu X D, et al. A novel fuzzy compensation multi-class support vector machines. Applied Intelligence, 2007, 27(1): 21-28

[5] Zhu M L, Chen S F, Liu X D. Sphere-structured support vector machines for multi-class pattern recognition. Lecture Notes in Computer Science, 2003, 2639: 589-593

[6] Wang J, Neskovic P, Cooper L N. Bayes classification based on minimum bounding spheres. Neurocomputing, 2007, 70: 801-808

[7] Schölkopf B, Burges C, Vapnik V N. Extracting support data for a given task // Proceedings of First International Conference on Knowledge Discovery and Data Mining. 1995: 252-257

[8] Schölkopf B, Platt J, Shawe-Taylor J, et al. Estimating the support of a high-dimensional distribution. Neural Computation, 2001, 13(7): 1443-1471

[9] Tax D, Duin R. Support vector domain description. Pattern Recognition Letter, 1999, 20: 1191-1199

[10] Tax D, Duin R. Support vector data description. Machine Learning, 2004, 54: 45-66

[11] Lee D, Lee J. Domain described support vector classifier for multi-classification problems. Pattern Recognition, 2007, 40: 41-51

[12] Song Q, Hu W J, Xie W F. Robust support vector machine with bullet hole image classification. IEEE Trans. on Systems, Man and Cybernetics, 2002, 32(4): 440-448

[13] Hu W J, Song Q. An accelerated decomposition algorithm for robust support vector machines. IEEE Trans. on Circuits and Systems, 2004, 51(5): 234-240

[14] Lin C F, Wang S D. Fuzzy support vector machines. IEEE Trans. on Neural Networks, 2002, 13(2): 464-471

[15] Lin C F, Wang S D. Training algorithms for Fuzzy support vector machines with noisy data // IEEE 8th Workshop on Neural Networks for Signal Processing. 2003: 517-526

[16] Krishnapuram R, Keller J M. A possibilistic approach to clustering. IEEE Trans. on Fuzzy System, 1993, 1(2): 98-110

[17] Krishnapuram R, Keller J M. The possibilistic c-means algorithm: insights and recommendations. IEEE Trans. on Fuzzy System, 1996, 4(3): 385-393

[18] Gardner A B, Krieger A M, Vachtsevanos G. One-class novelty detection for seizure analysis from intracranial EEG. Journal of Machine Learning Research, 2006(7): 1025-1044

[19] Lee S W, Lee S W. SVDD-based illumination compensation for face recognition // The 2nd International Conference on biometrics. LNCS 4642, 2007: 154-162

[20] Tao X M, Liu F R, Zhou T X. A novel approach to intrusion detection based on support vector data description // The 30th Annual Conference of IEEE industrial Electronics Society. 2004: 2016-2021

[21] http://www.ics.uci.edu/~mlearn/MLRepository.html

(上接第130页)

标$(x, y1)$或$(x, y2)$处在二值化图中为1，计数器n加1，设一个阈值k，如果$(n/(2r*2)) > k$，则椭圆存在，否则椭圆不存在。

4.3.3 *旋转*

结束语 本文介绍了印鉴识别系统中的印鉴录入过程，在本文中提到的精确查找圆章、方章、椭圆章的方法在实验中效果不错，并且已经应用到实际的印鉴识别系统中。

参 考 文 献

[1] 侯宇. 圆和椭圆边缘检测的快速方法. 中国计量学院学报, 2000

[2] 娄震，胡钟山，杨静宇. 支票自动处理系统中的图像处理及识别. 南京大学理工大学学报, 1999

[3] 求是科技. Visual C++数字图像识别技术典型案例. 北京：人民邮电出版社, 2004

[4] 王怀群. 二值图象的细化. 无锡轻工大学学报, 2001

[5] 屈稳太. 基于弦中点Hough变换的椭圆检测方法. 浙江大学学报(工学版), 2005